Reliability of an automatic classifier for brain enlarged perivascular spaces burden and comparison with human performance

Reliability of an automatic classifier for brain enlarged perivascular spaces burden and comparison with human performance

Víctor González-Castro[a,*], María del C. Valdés Hernández[a,*], Francesca M. Chappell[a], Paul A. Armitage[b], Stephen Makin[a], Joanna M. Wardlaw[a]

[a]*Department of Neuroimaging Sciences, Centre for Clinical Brain Sciences, University of Edinburgh, 49 Little France Crescent, Edinburgh EH16 4SB, UK*
[b]*Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield, Royal Hallamshire Hospital, Sheffield S10 2JF, United Kingdom*

Abstract

In the brain, enlarged perivascular spaces (PVS) relate to cerebral small vessel disease, poor cognition, inflammation and hypertension. We propose a fully automatic scheme that uses a support vector machine (SVM) to classify the burden of PVS in the basal ganglia (BG) region as low or high. We assess the performance of three different types of descriptors extracted from the BG region in T2-weighted MRI images: 1) statistics obtained from Wavelet transform's coefficients, 2) local binary patterns and 3) bag of visual words (BoW)-based descriptors characterising local keypoints obtained from a dense grid with the scale-invariant feature transform characteristics. When the latter were used, the SVM classifier achieved the best accuracy (81.16%). The output from the classifier using the BoW descriptors was compared with visual ratings done by an experienced neuroradiologist (Observer 1) and by a trained image analyst (Observer 2). The agreement and cross-correlation between the classifier and Observer 2 (κ=0.67[0.58 0.76]) were slightly higher than between the classifier and Observer 1 (κ=0.62[0.53 0.72]) and com-

*Corresponding Authors
Email addresses: victor.gonzalez@ed.ac.uk (Víctor González-Castro), M.Valdes-Hernan@ed.ac.uk (María del C. Valdés Hernández), F.Chappell@ed.ac.uk (Francesca M. Chappell), p.armitage@sheffield.ac.uk (Paul A. Armitage), Stephen.Makin@glasgow.ac.uk (Stephen Makin), joanna.wardlaw@ed.ac.uk (Joanna M. Wardlaw)

parable between both observers (κ=0.68[0.61 0.75]). Finally, three logistic regression models using clinical variables as independent variables and each of the PVS ratings as dependent variable were built to assess how clinically meaningful were the predictions of the classifier. The goodness-of-fit of the model for the classifier was good (AUC values 0.93(model1), 0.90(model2) and 0.92(model3)) and slightly better (i.e. AUC values 0.02 units higher) than that of the model for Observer 2. These results suggest that, although it can be improved, an automatic classifier to assess PVS burden from brain MRI can provide clinically meaningful results close to those from a trained observer.

Keywords: Brain MRI, Perivascular spaces, Discrete Wavelet transform, Local binary patterns, Bag of visual words, Support vector machine

1. Introduction

Perivascular spaces, also known as Virchow-Robin spaces, are fluid-containing spaces that surround the walls of small vessels and capillaries in the brain as they go through the grey or white matter. Perivascular spaces are microscopic, filled with interstitial fluid and act as drainage pathways for fluid and metabolic waste from the brain and, when enlarged, are visible in structural magnetic resonance imaging (MRI) sequences (Potter et al., 2015b). High number of enlarged perivascular spaces (PVS) has been reported to be associated with worse cognition (MacLullich, 2004), active inflammation in multiple sclerosis plaques (Wuerfel et al., 2008) or ageing (Aribisala et al., 2014), depression at older ages (Patankar et al., 2007), Parkinson's disease (Laitinen et al., 2000) and cerebral small vessel disease (Doubal et al., 2010).

The term Small Vessel Disease (SVD) refers to a group of pathological processes that affect the small arteries, veins and capillaries of the brain (Pantoni, 2010). It is the most common cause of vascular dementia and a cause of about a fifth of the strokes worldwide (Wardlaw et al., 2013), proven to have significant and strong associations with vascular risk factors (Staals et al., 2014).A moderate to severe burden of PVS in the basal ganglia (BG) is one of the markers of SVD (Wardlaw et al., 2013), along with lacunes, cerebral microbleeds and white matter hyperintensities (WMH).

PVS can be better identified on T2-weighted (T2w) MRI, where they appear as linear or dot-like structures with intensities close to those of the cerebrospinal fluid (CSF) and less than 3mm diameter in cross section (Wardlaw

et al., 2013). Therefore, PVS can be potentially quantified. Visual counting and/or manual delineation of PVS can be time consuming, and the development of computational methods to assess them is challenging, partly due to inconsistencies within the literature regarding PVS diameter and overlap in shape, intensity, location and size with these of lacunes (Valdés Hernández et al., 2013). Recently, Wang et al. (2016) and Ramirez et al. (2015) presented computational methods to obtain quantitative measurements of PVS and validated the usefulness of their procedures in clinical research, but both approaches are semi-automatic being, therefore, prone to inter-observer variations and could be time consuming. Cai et al. (2015) also proposed a method for quantifying PVS using high resolution 7T MRI scanners but the use of such field strengths, although providing good spatial resolution and signal-to-noise ratio, has limited clinical use. Ballerini et al. (2016) use a Frangi filter whose parameters are optimised by means of the ordered logit model to enhance the differentiation between PVS and the background, but is unsuitable for images with very anisotropic voxels commonly used in clinical settings (e.g. voxel sizes of 0.5 x 0.5 x 6 mm) and still requires the (visual) rating of the PVS.

As an alternative to quantitative measurements, several visual rating scales that provide a qualitative assessment of the burden of PVS have been proposed in recent years. Potter et al. reviewed the ambiguities of these scales and combined their strengths to develop one that proved to be robust (Potter et al., 2015a). However, as with any visual recognition process, it is subject to observer bias. Making the PVS rating automatic (e.g. replicating the visual rating scale using image processing and pattern recognition) could potentially overcome these and also the drawbacks that the current methods of PVS segmentation have.

Computer vision and pattern recognition have already been successfully applied for computer-aided diagnosis using MRI (Munsell et al., 2015; Beheshti and Demirel, 2015) and for segmentation of brain structures or lesions Ithapu et al. (2014); Roy et al. (2015); de Brebisson and Montana (2015). It has also been used to assess markers of SVD qualitatively. For example, Chen et al. proposed a framework based on multiple instance learning to distinguish between absent/mild vs. moderate/severe SVD in computed tomography (CT) scans (Chen et al., 2015).

However, to the best of our knowledge, only two papers have addressed the task of assessing automatically the PVS rating in brain MRI using computer vision and pattern recognition techniques (González-Castro et al., 2016;

González-Castro et al., 2016). They explored the use of different descriptors for this task, but did not analyse agreement with a human observer other than with the one that provided the ground truth ratings, or whether the predictors of the classification were clinically meaningful. Moreover, each of these two works evaluate different descriptors to characterise the brain region selected for classifying PVS burden and report similar levels of accuracy for the preferred schemes, albeit having validated the schemes differently (i.e.González-Castro et al. (2016) uses cross-validation and González-Castro et al. (2016) compares results on randomly divided train and test subsets). An overall evaluation of the schemes proposed so far for classifying the burden of PVS from brain MRI is lacking.

In this paper we build upon the work presented in (González-Castro et al., 2016; González-Castro et al., 2016), comparing the performance of the descriptors proposed by both studies for classifying automatically the burden of PVS using a Support Vector Machine (SVM) (Vapnik, 1995). We focus on the PVS in the basal ganglia (BG), since moderate to severe PVS in this region (i.e. ratings 2-4) is a marker of cerebral SVD. We evaluate three different types of descriptors: 1) statistics obtained from Wavelet transform's coefficients (Alegre et al., 2012), 2) local binary patterns (Ojala et al., 2002) and 3) bag of visual words (BoW)-based descriptors, using keypoints obtained from a dense grid characterised with the scale-invariant feature transform (SIFT) characteristics. Moreover, we validate the results by comparing the predictions made by the automatic method (i.e. the classifier using the descriptors that achieve the best performance) with the ratings from two observers. Finally, we also investigate the applicability of this classifier to clinical studies, to assess if its outcome is clinically meaningful. The paper is organised as follows: In Section 2 the dataset and proposed methods are explained. Section 3 introduces the experimental setup and the results of the experiments, which are discussed in Section 4. Finally, the conclusions and possible future lines of work are presented in Section 5.

2. Materials and methods

2.1. Subjects and MRI protocol

We used data from 264 patients who gave written informed consent to participate in a study of lacunar stroke mechanisms (Valdés Hernández et al., 2015).

The study that provided data for this manuscript (Valdés Hernández et al., 2015) included patients with lacunar stroke and/or minor cortical strokes which were clinically evident, and did not consider diabetes, hypertension and other vascular risk factors as criteria for exclusion. However, it excluded patients with other non-vascular neurological disorders, major medical conditions including renal failure, contraindications to MRI, unable to give consent, and those who had haemorrhagic stroke or whose symptoms resolved within 24 hours (i.e. transient ischaemic attack). It was approved by the Lothian Ethics of Medical Research Committee (REC 09/81101/54) and the NHS Lothian R+D Office (2009/W/NEU/14) and was conducted according to the principles expressed in the Declaration of Helsinki.

Brain MRI was conducted at baseline (i.e. there was a maximum of 8 days between the stroke and the scan) on a 1.5 tesla GE Signa LX clinical scanner (General Electric, Milwaukee, WI), equipped with a self-shielding gradient set and manufacturer supplied eight-channel-phased array heal coil. For our analyses we used the T2w images, acquired with TE 147 milliseconds, TR 9002 milliseconds, field of view 240×240 mm, acquisition matrix 256×256, slice thickness 5 mm, 1 mm inter-slice gap and voxel size $0.469 \times 0.469 \times 6$ mm. The reconstructed image size (in voxels) is $512 \times 512 \times 28$. For tissue segmentation, diffusion-weighted and structural T1-weighted (T1w), T2w and gradient echo, acquired as specified in (Valdés Hernández et al., 2015) were also used.

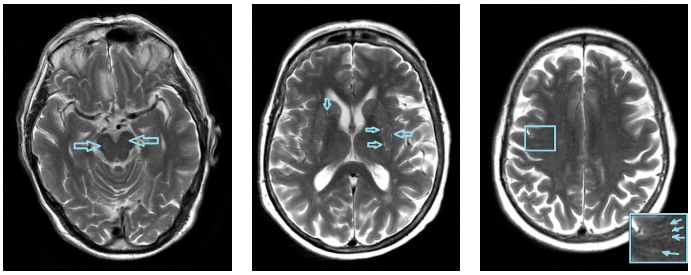
2.2. PVS visual rating scale

The visual rating scale proposed by Potter et al. was used for assessing the burden of PVS in the sample (Potter et al., 2015a). It rates the PVS separately in three major anatomical brain regions, i.e. midbrain, basal ganglia (BG) and centrum semiovale (CS) – shown in Figure 1 – using T2w MRI. The rating is done separately for left and right hemispheres, but a combined score that represents the average of the PVS burden is given.

In each of these anatomical regions, the rating can be 0 (no PVS), 1 (mild; 1-10 PVS), 2 (moderate; 11-20 PVS), 3 (frequent; 21-40 PVS) or 4 (severe; >40 PVS)[1].

All visual ratings were made by two observers: a neuroradiologist (Observer 1) with more than 25 years of experience who participated in the

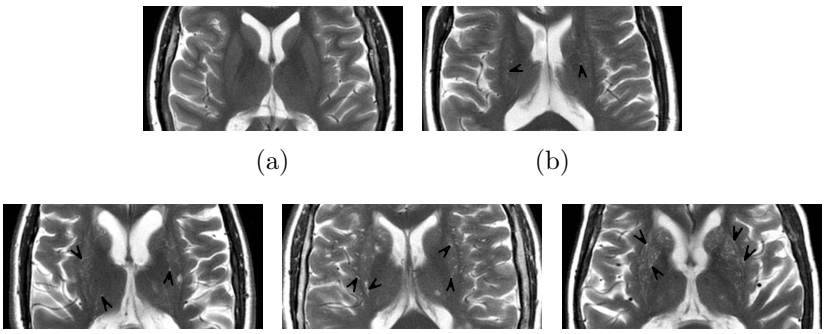
[1] `http://www.sbirc.ed.ac.uk/documents/epvs-rating-scale-user-guide.pdf`.



(a) (b) (c)

Figure 1: Example of the anatomical regions where the PVS (arrowed) are rated: Midbrain, Basal Ganglia and Centrum Semiovale (from (a) to (c), respectively).Note the longitudinal appearance in the centrum semiovale in axial view ((c)inset)

development of the scale and a trained image analyst (Observer 2). The ratings were done blind to all clinical information, each other's results and any intermediate or final computational results.



(a) (b)

(c) (d) (e)

Figure 2: Example for the PVS ratings in the BG, from 0 (none) to 4 (many) ((a) to (e), respectively) with black arrowheads pointing to some of the PVS.

In this paper, we focus only on the PVS in the BG, since moderate to severe PVS in this region (i.e. ratings 2-4) is a marker of cerebral SVD, which has been associated with cognitive decline (Staals et al., 2014), vascular dementia and stroke (Potter et al., 2015b). An example of each of the ratings for the BG is shown in Figure 2. We dichotomise the BG PVS scores into two classes as per Potter et al. (2015b), scores 0-1 (i.e. none or mild PVS burden) and scores 2-4 (i.e. moderate to severe), to be our classes 0 and 1,

respectively.

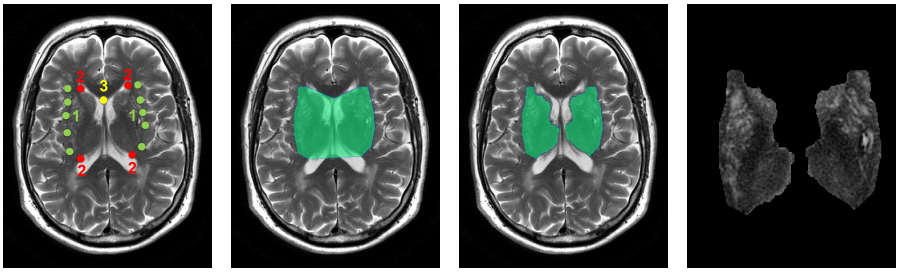
2.3. Image preprocessing

The guidelines for the visual rating of PVS according to this scale state that the rating should be done on the slice with the highest number of PVS, so as to minimize inconsistencies and intra-/inter-observer variations due to inter-slice variations in PVS visibility, varying number of PVS on different slices and double counting of linear PVS (Potter et al., 2015a). In the case of the BG region, this slice should be chosen amongst the slices with at least one characteristic BG structure, as indicated by Wang et al. (2016). A pipeline to extract the BG region and find the axial slice (from the BG) with the highest number of PVS for each subject, was developed.

The first step of this pipeline is to automatically segment the intracranial volume and cerebrospinal fluid (CSF) on the T1w images. This was achieved using optiBET (Lutkenhoff et al., 2014) and FSL-FAST (Zhang et al., 2001) respectively. The second step is to, also automatically, extract all subcortical structures, which was achieved using other tools from the same FMRIB Software Library (FSL) as is described in Valdés Hernández et al. (2015). Thereafter, from the slices that contained BG structures, we selected those in which the total area of these structures was more than 5 % the area of the intracranial area defined on the slice.

On each of the BG slices initially selected, a polygon enclosing the BG, internal and external capsules and thalami was automatically drawn by joining anatomical points in the insular cortex, the closest points to them in the lateral ventricles (frontal and occipital horns) and the intercept of the genu of the corpus callosum with the septum; and subtracting from it the region occupied by the CSF. These steps are illustrated in Figure 3.

From this subset of slices, the slice where our classifier operated was selected after applying contrast-limited adaptive histogram equalisation (CLAHE) (Zuiderveld, 1994) to the polygonal regions, thresholding them to 0.43 times the maximum intensity level (Valdés Hernández et al., 2013; Wang et al., 2016) (Fig. 3(d)), and counting the number of thresholded hyperintense regions on each candidate slice with area between 3 and 15 times the in-plane voxel dimensions (Wang et al., 2016). Although this procedure overestimates the number of PVS in the presence of other features of SVD markers (e.g. small lesions and lacunes) (Valdés Hernández et al., 2013), it provides a good estimate of the number of PVS on each candidate axial slice, so as to select the one with more PVS.



(a) (b) (c) (d)

Figure 3: Steps of the BG segmentation: (a) Detection of the vertices in the insular cortex (1), lateral ventricles (2) and genu (3); (b) creation of the polygon; (c) subtraction of the CSF from the BG polygonal region and (d) segmented BG region

2.4. Descriptors

2.4.1. Descriptors based on the Wavelet transform

The information represented by spatial frequencies has often been used for texture description with successful results (Arivazhagan and Ganesan, 2003). Due to its frequency domain localization capability, we have applied the discrete Wavelet transform (DWT) to each selected region to characterise their textures. We have used the Haar family of wavelets, which have already been successfully used in other medical image classification applications (Alegre et al., 2012). The DWT extracts the low and high frequency components of a signal so they can be analysed separately.

When the transform is applied to an image, four matrices of coefficients are obtained: namely LL_i, LH_i, HL_i and HH_i where i stands for the level of decomposition, which represent the approximations and details in the vertical, horizontal and diagonal directions respectively, They can be seen in the example that Figure 4 illustrates.

The first level of decomposition is applied on the original image, while the next levels i are applied to the matrix of approximations of level $i-1$ as Figure 5 shows.

One of the descriptors we used is based on the DWT, and it is built using the mean and standard deviations of the histograms of the original image and each one of the matrices of coefficients yielded after three DWT levels (i.e. LL_1, LH_1, HL_1, HH_1, LL_2, LH_2, HL_2, HH_2, LL_3, LH_3, HL_3 and HH_3). Hence we represent each region by a vector of 26 features. This descriptor is known as Wavelet statistical features (WSF) (Arivazhagan and Ganesan,

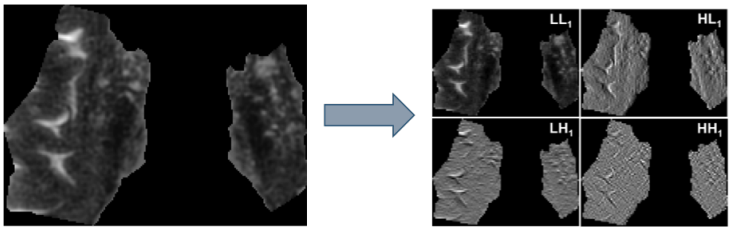


Figure 4: First level DWT decomposition of the basal ganglia region from a T2w image.

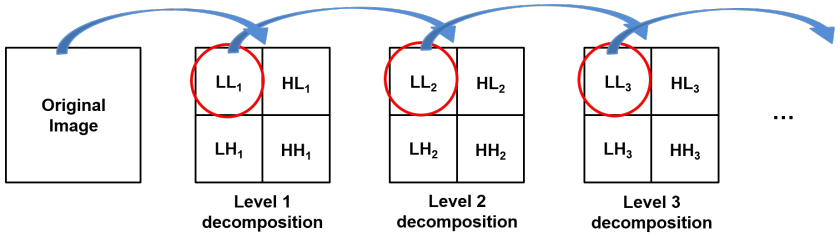


Figure 5: Example of the names of the coefficient matrices after a three-level DWT decomposition.

2003; Alegre et al., 2012).

The other descriptor based on the DWT is built using the features proposed by Haralick et al. (1973) derived from the grey-level co-occurrence matrix (GLCM) of the original image and each of the the coefficient matrices obtained after the first DWT level (i.e. LL_1, LH_1, HL_1 and HH_1). The features extracted from each GLCM are concatenated to form the final descriptor. A diagram depicting this process is shown in Figure 6.

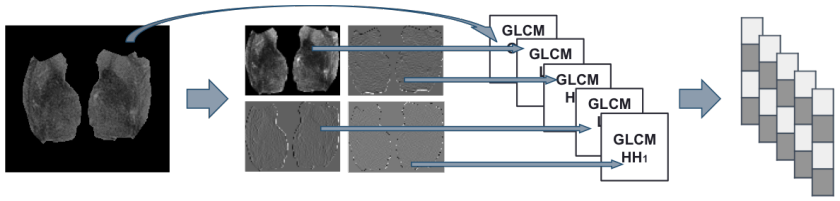


Figure 6: Diagram showing how the WCF descriptors are built.

To achieve some invariance to rotation, we averaged the features extracted from GLCMs computed with orientations 0°, 45°, 90° and 135°. These descriptors are called Wavelet Co-occurrence Features (WCF) (Arivazhagan and Ganesan, 2003; Alegre et al., 2012). In this work, we assess two vari-

ants of the WCF descriptors, WCF_4 and WCF_{13}, depending on whether we extracted 4 or 13 features from the GLCMs, respectively. WCF_4 is built using the Haralick features *Contrast*, *Correlation*, *Energy* and *Homogeneity*, and WCF_{13} is formed using all features proposed by Haralick et al. (1973) except the *Maximal Correlation Coefficient*. These two descriptors showed good performance in Alegre et al. (2009).

2.4.2. Local binary patterns

Local Binary Patterns (LBP) were introduced by Ojala et al. (2002). In the original version they worked with a 3×3 pixel block, but LBPs were later generalised, so as the size of the neighbourhood and the number of sampling points were parameters of the method. Given a pixel c with coordinates (x_c, y_c), a pattern code is calculated by comparing it with the value of its P neighbours separated by a distance R, which in our case is 1, as per Equation (1).

$$LBP_{R,P} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p, \tag{1}$$

where g_c and g_p are the grey-level values of pixel c and its p-th neighbour, and function $s(g_p - g_c)$ is defined as:

$$s(g_p - g_c) = \begin{cases} 1 & \text{if } g_p - g_c \geq 0 \\ 0 & \text{if } g_p - g_c < 0 \end{cases}$$

Finally, the whole image is described by means of a histogram of the LBP values of all pixels, given by Equation (1). As the position of the *first* neighbour (i.e. $p = 0$) is fixed, it being the pixel on the right hand side of c, the $\text{LBP}_{R,P}$ operator is not invariant to rotation We remove such effect of rotation using the rotation invariant local binary pattern, $LBP^{ri}_{R,P}$, defined in Ojala et al. (2002).

As certain local binary patterns represent fundamental properties of texture, providing the vast majority of patterns present in textures (Ojala et al., 2002), while others are known to be less descriptive of the texture, Ojala et al. introduced a measure of 'uniformity' $U(LBP_{R,P})$, which counts the number of spatial transitions (i.e. bitwise 0/1 changes) in a binary pattern $\text{LBP}_{R,P}$ for $\text{LBP}_{R,P}$ less than 2 (i.e. $\text{LBP}^{riu2}_{R,P}$) as expressed in Equation (2).

$$LBP^{riu2}_{R,P} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c) & \text{if } U(LBP_{R,P}) \leq 2 \\ P + 1 & \text{otherwise} \end{cases}, \tag{2}$$

As the BG regions and the PVS are not very big we tried to keep the texture analysis as local as possible, so in this work we have used the values $R = 1$ and $P = 8$. The final descriptors we use are the histograms of the accumulated output of $\text{LBP}_{1,8}$, $\text{LBP}^{ri}_{1,8}$ and $\text{LBP}^{riu2}_{1,8}$ operating in each BG region.

2.4.3. Bag of visual words

The Bag of Visual Words (BoW) model (Sivic and Zisserman, 2003) represents each image as a function of the frequency of appearance of certain visual elements, called visual words. The set of visual words is called the *dictionary* or *codebook*.

To build the dictionary, a set of keypoints from each image are sampled. Around each keypoint a small square region (i.e. patch) is extracted and characterised by means of descriptors that retrieve information about the distribution of its pixels intensities. After that, the descriptors of the patches are clustered into K groups, each one having a prototype feature vector which is called *visual word*. This process is depicted in Figure 7.

In this work, we use a dense grid for sampling the keypoints and the k-means clustering method (MacQueen, 1967) for forming the visual words. The process of creating the dictionary is performed in each iteration of the cross validation using the subsets of images used for training. We assessed different numbers of visual words to evaluate their impact on the classification.

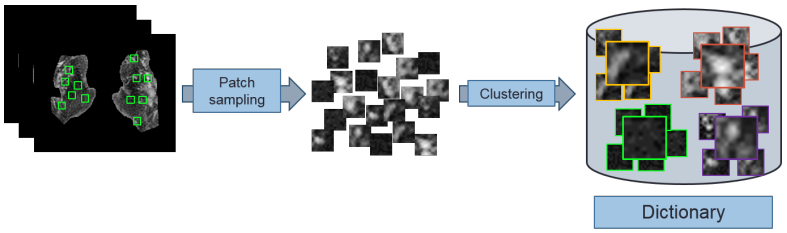


Figure 7: Diagram showing how the dictionary is created.

Once the dictionary is built, each image of the dataset is described by means of a process called *image representation*. This consists of repeating, for each image, the same process of keypoint selection and characterisation used in the creation of the dictionary, using also the same methods. Then, for each "new" patch, we find the visual word of the dictionary that is most

similar to it by means of calculating the Euclidean distance between their descriptors.

The histogram of the visual words representative of all patches in an image is used as its final descriptor. The image representation process is illustrated in Figure 8.

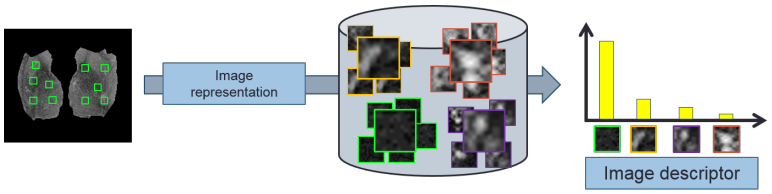


Figure 8: Diagram showing how the image representation is carried out

In this work, the patches are described using the Scale Invariant Feature Transform (SIFT) (Sivic and Zisserman, 2003). Basically, SIFT descriptors are based on histograms of oriented gradients computed from the intensities of the regions that result from dividing a 16×16 pixel squared patch around each keypoint into 16 subregions of 4×4 pixels each. More details about SIFT can be found in Sivic and Zisserman (2003). Despite these consisting of two different parts, keypoint detector and patch descriptor, we only use the patch descriptor as we are sampling the keypoints in a dense grid.

2.5. Classification

In this work, we use a Support Vector Machine (SVM) classifier, which is a supervised machine-learning approach that adjusts internal "weights" by means of a training process (i.e. an optimization phase), minimising the error between its calculated response and a "ground truth" provided by an expert. This type of classifier has attracted attention in the last few years for analysing MR images (Nam et al., 2015; Tong et al., 2014; Feis et al., 2013). SVM tries to find the optimal hyperplane that maximizes the distances (i.e. margins) to the instances of the positive and negative classes in the training dataset. One of the parameters of SVM is the cost parameter C, which controls the trade-off between classes allowing training errors and forcing rigid margins.

SVM is a linear classifier: it tries to separate the data using a linear hyperplane. There are cases where the data is not linearly separable. In those cases, SVM may use the *kernel trick*: A kernel function $\mathrm{K}(\mathbf{x}', \mathbf{x})$ may

transform the data into a higher dimensional space where it is possible to separate it linearly. After evaluating different kernels (i.e. linear, radial basis function, sigmoid), the best results were achieved with the radial basis function (RBF) kernel:

$$\mathrm{K}(\mathbf{x}', \mathbf{x}) = \exp(-\gamma \left\|\mathbf{x}' - \mathbf{x}\right\|^2) \tag{3}$$

We refer the reader interested in more details about SVM to Schölkopf and Smola (2001).

We use several combinations of the regularization parameter C (i.e. 1, 5, 10, 50, 100, 250 and 500) and γ (i.e. 10^{-5}, 10^{-4}, 10^{-3}, 0.01 and 1), assessed with all descriptors, to find the optimal configuration. We use the implementation provided in the libSVM library[1] (Chang and Lin, 2011).

2.6. Validation of the classifier

We validated the classification with a stratified 5-fold cross validation as follows. The whole set, represented by the descriptors explained in Sections 2.4.1, 2.4.2 and 2.4.3, was randomly partitioned into 5 equally sized subsets with the same distribution as the original set. Of the 5 subsets, 4 were used to train the classifier and the remaining one was used as the test set. This process was repeated 5 times using a different subset each time as test set. The 5 results from the 5 folds were averaged to provide the final results.

This cross validation process was repeated 10 times, and the 10 results were averaged to avoid possible bias due to a random separation of the folds. Data were normalised so that they had mean 0 and standard deviation 1.

The overall results were validated in terms of accuracy, sensitivity and specificity, using the dichotomised ratings of Observer 1 as ground truth.

2.7. Statistical analyses

The descriptors that achieve the best performance would be used in a real automatic visual rating application. Therefore we analysed the agreement of the visual ratings between the automatic classifier based on those descriptors and between each observer. We also analysed the association between the outcome of each PVS rating (i.e. from each observer and from the automatic classifier) and clinical parameters known to be related to PVS burden in the patients that comprise this sample (see Section 2.1).

[1] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

2.7.1. Inter-observer agreement

We determined the weighted Kappa coefficient of the PVS ratings in the BG region (scale 0-4) between observers as per `http://vassarstats.net/kappa.html` (Copyright Richard Lowry 2001-2015). We also performed marginal homogeneity tests of the basal ganglia PVS visual ratings (scale 0-4) using the software application mh.exe ver. 1.2 (2016-03-01) (by John Uebersax).

After dichotomising the BG PVS visual ratings produced by both observers, we determined the Kappa coefficient between observers and between the automatic classifier and each observer, using the function kappa in MATLAB R2015a (Copyright (c) 2007, Giuseppe Cardillo, updated 23 Dec 2009, `http://uk.mathworks.com/matlabcentral/fileexchange/15365-cohen-s-kappa/content/kappa.m`). We also conducted the McNemar's test between the ratings produced by the expert (i.e. Observer 1) and the automatic classifier to investigate whether the marginal frequencies between both were or not equal.

2.7.2. Clinical validation

The following clinical and demographic parameters were available for each study participant: age, hypertensive (or not) classification, stroke subtype (lacunar or cortical) classification and scores of white matter hyperintensity (WMH), atrophy and SVD burden. WMH were coded using Fazekas scores, for periventricular (PV) and deep lesions separately in the left and right hemispheres and a combined score for both hemispheres was recorded (Fazekas et al., 1987). Brain atrophy was coded using a validated age-relevant template (Farrell et al., 2008), with superficial and deep atrophy coded separately ranging from none to severe on a scale from 1 to 6 according to the centiles into which the template is divided, being 1($< 25^{th}$), 2(25-50^{th}), 3(50-75^{th}), 4(75-95^{th}), 5($> 95^{th}$) and if >>5, 6 is used. Total atrophy was calculated as the average of deep and superficial atrophy scores. SVD was coded as per Staals et al. (2015) (0-4), which confers a point for each of the following conditions: if 1 or more cavitated old lacunar lesions are present, if Fazekas PV score >= 3 and/or Fazekas Deep score >= 2, if BG PVS score is >= 2 as per Potter et al. (i.e. moderate-to-extensive), and if more than 1 brain microbleed is present.

We calculated the non-parametric bootstrapped correlations between BG PVS scores (before and after dichotomisation, from observers and from the automatic classifier) and each clinical variable. We also performed bino-

mial multivariable logistic regression to evaluate the clinical usefulness of our machine-learning scheme as per Potter et al. (2015a) and its sensitivity in various models. The latter was evaluated by comparison of correlated receiver operating characteristic (ROC) curves obtained from three models that have as outcome variable the dichotomised PVS rating from A) the automatic classifier, B) Observer 1, and C) Observer 2. The first model (i.e. Model 1) had the following predictors: age, total atrophy, hypertension, Fazekas score, whether the patient had a previous lacunar infarct or not, index stroke subtype and SVD score. The second model (i.e. Model 2, implemented in Potter et al. (2015a)) had the same predictors as Model 1 with the exception of SVD score. The third model (i.e. Model 3) had also the predictors of Model 1 with the exception of Fazekas score and whether the patient had a previous lacunar infarct or not, as these two parameters are contemplated within the SVD score. These analyses were done using MATLAB R2015a. Of note, the PVS outcome variable is also a contributor to the SVD score.

2.7.3. Analysis of the robustness against imaging confounds

All scans of the primary study that provided data for this analysis underwent quality checks. None of the T2-weighted sequences were corrupted by visible movement artefacts that could affect the automatic PVS rating procedure presented. However, there are other confounds that could have influence in the results. We calculated the number of scans misclassified on each of the 10 iterations that contributed to the final result, on the absence and presence of the following imaging confounds visually identified by Observer 2 in the basal ganglia region blind to the neuroradiological reports: white matter hyperintensities found either bilaterally and scattered throughout the region or as a single cluster possibly indicative of a recent or old subcortical infarct, lacunes (symptomatic or asymptomatic), recent or old cortical strokes that partially affect the region, globus pallidus partially or totally hyperintense, partial volume effects of the cerebrospinal fluid, and a combination of two or more of these factors.

We also counted the number of scans misclassified on each iteration for those people who had a lacunar infarct neuroradiologically determined, regardless of whether it was visible on T2-weighted in the basal ganglia region or not. This analysis would allow us to discuss whether the occurence of a recent lacunar infarct influenced the descriptors used by the classifier.

3. Results

The PVS ratings made by the experienced neuroradiologist (Observer 1), used to train the classifier, were distributed across the sample as Table 1 shows. The dichotomisation of these ratings into none-mild vs. moderate-severe resulted in 133 and 131 datasets for each class, respectively.

Table 1: Distribution of the visual ratings in the sample.

PVS rating	**0**	**1**	**2**	**3**	**4**	**TOTAL**
Num. images (%)	5 (1.89%)	128 (48.48%)	68 (25.76%)	44 (16.67%)	19 (7.20%)	264

3.1. Results of the SVM classification

Table 2 shows the best results using the descriptors based on the Wavelet transform (i.e. WSF, WCF_4 and WCF_{13}), the descriptors based on local binary patterns with $R = 1$ and $P = 8$ (i.e. $\mathrm{LBP}_{1,8}$, $\mathrm{LBP}^{ri}_{1,8}$ and $\mathrm{LBP}^{riu2}_{1,8}$), the fusions of the descriptors WCF_4 and WCF_{13} with $\mathrm{LBP}^{riu2}_{1,8}$ and the descriptors based on the bag of visual words model.

Table 2: Average accuracy (acc.), sensitivity (sens.) and specificity (spec.), as well as their respective standard deviations of the SVM 5-fold classification along the 10 iterations. Also, the parameters C and γ theses results were obtained with are provided.

	C	γ	**Acc. (%)**	**Sens. (%)**	**Spec. (%)**	**std**	**std**$_{sens}$	**std**$_{spec}$
WSF	500	10^{-4}	73.47	78.71	68.14	0.90	1.13	1.26
WCF_4	50	10^{-4}	73.66	77.15	70.12	1.25	1.38	1.68
WCF_{13}	250	10^{-4}	75.95	77.86	73.96	0.87	1.63	1.04
$\mathrm{LBP}_{1,8}$	50	10^{-3}	68.34	70.49	66.21	2.54	2.18	3.51
$\mathrm{LBP}^{ri}_{1,8}$	50	10^{-4}	70.02	75.95	64.01	1.02	1.22	1.49
$\mathrm{LBP}^{riu2}_{1,8}$	10	0.01	74.22	81.97	66.37	0.70	1.39	1.57
$\mathrm{WCF}_4 + \mathrm{LBP}^{riu2}_{1,8}$	250	10^{-4}	78.84	79.84	77.80	1.12	1.60	1.25
$\mathrm{WCF}_{13} + \mathrm{LBP}^{riu2}_{1,8}$	100	10^{-4}	78.13	78.62	77.58	1.16	2.07	1.55
BoW	5	10^{-4}	81.16	79.31	82.97	1.72	2.20	2.57

The best descriptor in terms of overall accuracy was the descriptor based on the Bag of Visual Words model (81.15%) using a dictionary with 175 visual words, followed by the fusion of WCF_4 and $\mathrm{LBP}^{riu2}_{1,8}$ (78.84%). Moreover, the former reached a sensitivity just slightly worse than the latter. The highest sensitivity is achieved by $\mathrm{LBP}^{riu2}_{1,8}$, but its specificity is much worse than the BoW-based descriptor. It is also remarkable that, whereas WCF_4 does not

get a good accuracy on its own, its accuracy improves 7% when it is fused with the $\text{LBP}^{riu2}_{1,8}$ descriptor.

The automatic classifier used in the following sections will be the SVM based on the descriptors that achieved the best overall accuracy (i.e. the dense-SIFT-based Bag of Visual Words model, with the SVM parameters $C = 5$ and $\gamma = 10^{-4}$ using a dictionary of 175 visual words). Once the visual dictionary is created and the classifier is trained, this method took 0.0477 seconds to describe and classify each image.

3.2. Inter-observer variability

The agreement of the BG PVS ratings (scale 0-4) between observers 1 and 2 was kappa = 0.8269, std. error 0.0398, 95% CI[0.749 0.9048]. The maximum possible linear-weighted kappa, given the observed marginal frequencies was 0.8729. McNemar's tests for each rating (0-4), and McNemar tests of equal thresholds were significant in rating 1 ($p<0.003$).

The agreement of the dichotomised BG PVS ratings between observers 1 and 2 was kappa = 0.6822, std. error 0.0369 and 95% CI[0.6099 0.7545]. The maximum possible linear-weighted kappa, given the observed marginal frequencies was 0.8486.

Table 3 shows the agreement (i.e. kappa coefficient, standard error, 95% CI and maximum possible linear-weighted kappa, given the observed marginal frequencies) between each observer and the ratings assigned by the SVM classifier that yield the best accuracy (see Table 2). Since the classification experiment was repeated 10 times, the reported agreements are the average of the corresponding 10 agreements. The marginal proportions between the ratings from the expert (i.e. Observer 1) and the automatic classifier were non-significantly different from each other (McNemar's test p=0.1086). See the 2x2 frequency Table 4.

Table 3: Kappa coefficient, standard error and 95% CI, given the observed marginal frequencies between each observer and the automatic classification method.

	Kappa	**std. Error**	**95% CI**
Obs. 1 vs. Classifier	0.6228	0.0481	$[0.5286, 0.7170]$
Obs. 2 vs. Classifier	0.6743	0.0455	$[0.5851, 0.7635]$

Table 4: Two-by-two table between the ratings done by the expert (i.e. Observer 1), the predictions of the classifier and ratings from Observer 2.

	Auto. Classifier		**Observer 2**	
Ratings	**0**	**1**	**0**	**1**
Observer 1, rating 0	104	29	102	31
Observer 1, rating 1	18	113	11	120

3.3. Clinical validation

3.3.1. Bootstrapped correlations between the PVS ratings and with the clinical parameters

Visual ratings done by Observer 1 (dichotomised and not dichotomised), Observer 2 (dichotomised and not dichotomised) and the automatic classifier were equally significantly and positively correlated with age, PVS ratings in centrum semiovale (dichotomised and not), atrophy (deep and superficial), Fazekas (deep and periventricular), hypertension, old lacunar infarcts and SVD score. None of the BG PVS ratings correlated with index stroke subtype (lacunar or cortical), and all were highly and significantly correlated with each other as Table 5 shows.

Table 5: Non-parametric bootstrapped cross-correlation matrix for PVS ratings in the basal ganglia region. All correlations shown were significant with $p<0.0001$.

Parameter	Observer 1 Scale 0-4	Observer 1 Dichotomised	Observer 2 Scale 0-4	Observer 2 Dichotomised	Automatic Classifier
Observer 1 (0-4)	1	0.9317	0.8130	0.6828	0.6588
Observer 1 (0-1)		1	0.7341	0.6901	0.6464
Observer 2 (0-4)			1	0.9057	0.7127
Observer 2 (0-1)				1	0.7030

3.3.2. Applicability in clinical research

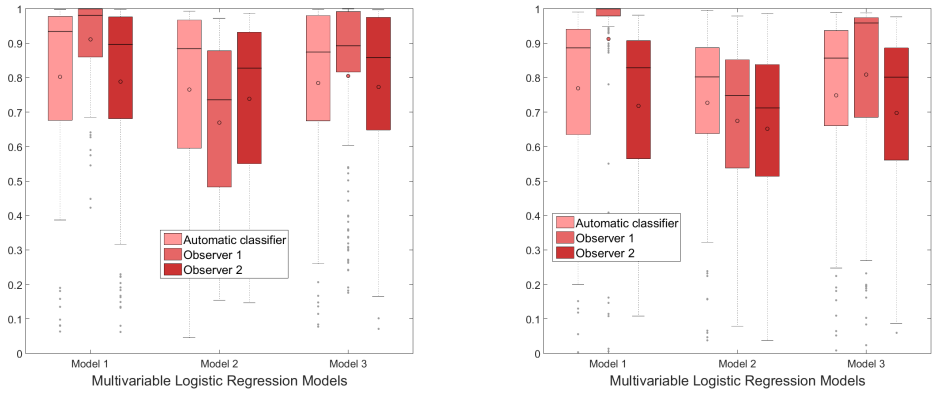
Table 6 shows the results of the binomial multivariable logistic regression. Age, Fazekas periventricular scores and the presence of old lacunar infarcts were significant and negatively associated with all BG PVS scores (i.e. those done by both observers and by the automatic classifier), as in Potter et al. (2015a). The coefficient estimates tabulated (B) express the effects of each predictor variable on the log odds of being in one class (i.e. 1 or 0) versus the reference class (i.e. 1 or 0 as per Observer 1).

Table 6: Coefficient estimates and significance (B (p-value)) of the associations for each predictor (i.e. clinical parameter) for Model 2. The outcome variable is the dichotomised PVS score

Predictor	Automatic Classifier	Observer 1 Dichotomised	Observer 2 Dichotomised
Age (years)	0.0569 (0.0044)*	0.0462 (0.0074)*	0.0613 (0.0009)*
Atrophy (scale 1-6)	0.0075 (0.9313)	-0.0411 (0.5726)	-0.0472 (0.5545)
Hypertension (0-1)	0.2930 (0.4649)	-0.2131 (0.5534)	0.3675 (0.3120)
Fazekas Deep (0-3)	0.5394 (0.0874)	0.1231 (0.6570)	-0.0801 (0.7854)
Fazekas PV (0-3)	1.4615 (<0.0001)**	1.1553 (0.00012)*	1.3928 (<0.0001)**
Old lacunar infarcts (0-1)	0.9625 (0.0245)*	1.0139 (0.0063)*	1.1987 (0.0035)*
Index stroke lacunar (0-1)	0.2953 (0.4355)	0.4294 (0.1881)	0.1853 (0.5868)

3.3.3. Sensitivity analysis

Figure 9 shows the predicted probabilities of the outcome variables for each model. The distribution of the predicted "0"s and "1"s to be 0 and 1 respectively for the classifier and Observer 2 were similar across all models. All outcomes (i.e. PVS ratings from the classifier, Observer 1 and Observer 2) were consistently poorer for Model 2, which does not include SVD scores as predictor, than for the other two models. The PVS ratings from Observer 1 were particularly sensitive to the presence -and absence- of the SVD scores as predictor in the model, being exceptionally high when more components of the SVD score (including it) were included (i.e. Model 1).



(a) (b)

Figure 9: Boxplots showing the distributions of the predicted probabilities of the outcome variable "1" (a) and "0" (b) (i.e. PVS ratings from the automatic classifier, from Observer 1 or from Observer 2) for each logistic regression model.

Figure 10 shows the correlated ROC curves for each outcome variable (i.e. automatic classifier, Observers 1 and 2) also for each model. The area under the curve (AUC) from the automatic classifier experiences the least variation across the three curves: 0.93,0.90 and 0.92 for models 1,2 and 3 respectively (maximum variation 3%) indicating highest consistency in model accuracy, followed by Observer 2 (maximum variation 5%).

3.3.4. Performance on the presence/absence of imaging confounds

As table 7 shows, only 9.6% to 16.6% of the scans that have a small T2-weighted hyperintense lesion such as lacunes, white matter hyperintensities or subcortical new or old infarcts in the basal ganglia region of size comparable with those of the PVS were misclassified, versus 16% of the scans that have two or more of these confounds, and 13.6% of those who had none. These percentages were higher when the T2-weighted hyperintense covered a larger region (i.e. cortical stroke or globus pallidus hyperintense), but the number of scans that had these confounds were very small (7 and 5 respectively out of 264). The number of patients who had a recent lacunar infarct (neuroradiologically determined) and for which the PVS rating was miscalculated was the same as the number of patients that did not have any imaging confound and for which the PVS rating done by the classifier was wrong (compared to the ratings of the neuroradiologist).

Table 7: Number of scans misclassified per number of iterations, on the presence/absence of imaging confounds.

Confounds	Total no. scans	No. of scans misclassified 1-3 iter.	4-6 iter.	7-10 iter.
unilateral WMH	31	4	1	3
lacunes (symptomatic or not)	25	4	4	3
bilateral WMH	30	5	0	5
cortical stroke/CSF partial vol.	7	0	0	2
globus pallidus hyperintense	5	1	1	1
two or more of the above	56	4	3	9
none	110	12	8	15
lacunar stroke	119	17	5	15

4. Discussion

We developed an automatic framework to classify T2-weighted MRI as having none or few PVS in the basal ganglia region versus having many of them, in response to the need for such tool given the role of PVS in SVD and vascular dementia progression. Our framework uses a conventional SVM classifier based on the information from SIFT descriptors that operate on patches from the basal ganglia region using a dense grid following the "bag of words" model. These descriptors provided the highest classification accuracy (81.16%) from those evaluated. This accuracy is slightly lower than the one reported in González-Castro et al. (2016) with the same descriptors (82.34%). The reason is the different validation of the classifier used in both works: in González-Castro et al. (2016) the classification was carried out by randomly splitting the dataset into train (70%) and test sets (30%), whereas in this case we have used 5-fold cross validation. This classifier took an average of 0.0477 seconds to describe and classify each image. The framework proved to be useful in clinical settings and outperformed the visual classification done by a trained observer.

The image processing pipeline that pre-processed the data where the descriptors were extracted was designed following the visual rating guidelines for PVS from Potter et al (Potter et al., 2015a) (`http://www.sbirc.ed.ac.uk/documents/epvs-rating-scale-user-guide.pdf`), which are based on assessing the PVS from a region of interest on the axial MRI slice with the most visible PVS. All agreements between the automatic classifier, the

dichotomised ratings of the experienced neuroradiologist (Observer 1) and those from the trained observer (Observer 2), as shown in Section 3.2 were above 0.6. However, the agreement between the dichotomised ratings from both observers (kappa = 0.6822) was slightly higher than the agreement between the classifier and any of the observers (0.6228 with Observer 1 and 0.6743 with Observer 2). The fact that the classifier had better agreement with Observer 2 than with Observer 1 may be because Observer 2 followed the same guidelines used to design the pipeline for the automatic classifier, whereas Observer 1 may have also applied their individual experience and neuroradiological knowledge when rating the PVS. The cross-correlation between the classifier output and the dichotomised ratings of both observers, shown in Table 5, followed the same pattern: the correlation of the classifier with Observer 2 was higher than with Observer 1 (0.7030 and 0.6464, respectively). This cross-correlation between the output of the classifier and the dichotomised ratings of Observer 2 (0.7030) was comparable and even slightly higher than between the dichotomised ratings of both observers (0.6901).

The statistical model built to evaluate the applicability of the automatic classifier to the clinical research showed excellent and similar goodness-of-fit irrespective of whether the outcome variable was the automatic classifier (AUC=0.90), Observer 1 (AUC=0.84) or Observer 2 (AUC=0.86). Also, age, the burden of periventricular white matter hyperintensities (i.e. Fazekas PV) and the presence of old lacunar infarcts were associated with the PVS burden irrespective of whether these were rated automatically or visually by any of the observers, proving the usefulness of the automatic framework proposed. A separate sensitivity analysis of this and similar correlated models showed that the automatic classifier was the least susceptible to be influenced by the overall burden of SVD shown in the MRI scan whilst the ratings from the neuroradiologist captured better the full flavour of the SVD features. The degree in which this result was favoured by the single-slice approach adopted by the classifier (Potter et al., 2015a)(Wang et al., 2016) is not known. Further evaluation on the whole extent of the three anatomical regions defined by Potter et al. (2015a), with added scrutiny to exclude lacunes is needed. Nevertheless, given that the accuracy of the classifier on the presence of imaging confounds was not different from it in the absence of them, and that the output was quite robust against the whole SVD burden, we do not foresee any problem for this automatic classification scheme to be applied to longitudinal or multicentre studies, as long as the training and testing datasets have similar acquisition protocols.

A possible limitation of this work is the fact that the segmentation of the basal ganglia region is not always accurate (due to, for example, not finding the anatomical points described in Section 2.3), causing a potential misclassification. As we wanted to assess the validity of a fully automatic method, we kept those suboptimal segmentations. Another limitation of the study may be the dichotomisation of the visual ratings used in the automatic classification. Due to limitations in the sample size, we needed to simplify the classification, so we dichotomised the visual rating scale as it was done in previous studies (Potter et al., 2015b): a reliable 5-class classification model is not possible to be trained with such few instances in some classes (e.g. out of 264 subjects there were only 5 with rating 0 or 19 with rating 4). Further analyses using bigger samples and considering the full ratings (i.e. 0-4) need to be done.

5. Conclusions and Future Work

In this paper we have proposed an automatic framework based on image analysis and machine learning to predict the burden of enlarged perivascular spaces on the basal ganglia as "none or few" or "moderate to severe" based on the PVS visual rating scale Potter et al. (2015a).We compared different descriptors computed from the basal ganglia region. The bag-of-visual-words-based descriptors achieved the best accuracy (81.16%) in the classification, carried out using a support vector machine trained using the visual ratings provided by an experienced neuroradiologist (i.e., Observer 1) as ground truth.

We also compared the predictions of the classifier with the visual ratings done by Observer 1 and also with those done by a trained image analyst (i.e., Observer 2). The inter-observer agreement with the Observer 2 (kappa=0.6743) was higher than with the Observer 1 (kappa=0.6228) and comparable to that between both observers (kappa=0.6822). The cross-correlation with the Observer 2 (0.7030) is also higher than with the Observer 1 (0.6464), and slightly higher that that between both observers (0.6901).

Finally, we built three correlated logistic regression models with some clinical variables as independent variables and the ratings predicted by the automatic method and both observers as outcome variables and demonstrated that, although the automatic classifier does not capture the overall SVD severity, it can be used in clinical research as it consistently gives a meaningful output in relation to clinical parameters.

For future work we will try to improve the classification performance by means of extracting the whole basal ganglia region and use the information from all slices where the extracted region appears (i.e. 3D analysis), as it may provide information that we are currently not taking into account. We will also try to use data from patients from other studies to increase our sample size and perform a 5-class classification (i.e. ratings from 0-4). Supervised machine-learning schemes like the one presented here would require the ground truth PVS counts or segmentations from a large number of datasets done by an expert to be able to count and/or segment PVS. Such data are currently unavailable. However, the output from this classifier could be used as input to the fully automatic PVS unsupervised segmentation approach developed by Ballerini et al. (2016), (mentioned in the Introduction Section) which needs the PVS ratings to tune its algorithm and make it fully automatic. Finally, the classifier presented here could be adapted to get the visual rating of the PVS in the centrum semiovale.

Acknowledgements

We would like to thank study participants, radiographers and staff at the Brain Research Imaging Centre Edinburgh, a SINAPSE (Scottish Imaging Network A Platform for Scientific Excellence) collaboration centre, the Wellcome Trust for funding the primary study that provided the data (Ref. No. 088134/Z/09) and the Row Fogo Charitable Trust (Grant No. BRO-D.FID3668413).

References

Alegre, E., González-Castro, V., Alaiz-Rodríguez, R., García-Ordás, M. T., 2012. Texture and moments-based classification of the acrosome integrity of boar spermatozoa images. Comput Methods Programs Biomed 108 (2), 873–881.

Alegre, E., González-Castro, V., Suárez, S., Castejón, M., Sept 2009. Comparison of supervised and unsupervised methods to classify boar acrosomes using texture descriptors. In: ELMAR, 2009. ELMAR '09. International Symposium. pp. 65–70.

Aribisala, B. S., Wiseman, S., Morris, Z., Valdes-Hernandez, M. C., Royle, N. A., Maniega, S. M., Gow, A. J., Corley, J., Bastin, M. E., Starr, J.,

Deary, I. J., Wardlaw, J. M., jan 2014. Circulating inflammatory markers are associated with magnetic resonance imaging-visible perivascular spaces but not directly with white matter hyperintensities. Stroke 45 (2), 605–607.

Arivazhagan, S., Ganesan, L., 2003. Texture classification using wavelet transform. Pattern Recognition Letters 24 (9?10), 1513–1521.

Ballerini, L., Lovreglio, R., del C. Valds Hernndez, M., Gonzalez-Castro, V., Maniega, S. M., Pellegrini, E., Bastin, M. E., Deary, I. J., Wardlaw, J. M., 2016. Application of the Ordered Logit Model to Optimising Frangi Filter Parameters for Segmentation of Perivascular Spaces. Procedia Computer Science 90, 61 – 67, 20th Conference on Medical Image Understanding and Analysis (MIUA 2016).

Beheshti, I., Demirel, H., 2015. Probability distribution function-based classification of structural MRI for the detection of Alzheimer's disease. Computers in Biology and Medicine 64, 208 – 216.

Cai, K., Tain, R., Das, S., Damen, F. C., Sui, Y., Valyi-Nagy, T., Elliott, M. A., Zhou, X. J., 2015. The feasibility of quantitative MRI of perivascular spaces at 7T. Journal of Neuroscience Methods 256, 151–156.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chen, L., Tong, T., Ho, C. P., Patel, R., Cohen, D., Dawson, A. C., Halse, O., Geraghty, O., Rinne, P. E., White, C. J., et al., 2015. Identification of Cerebral Small Vessel Disease Using Multiple Instance Learning. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer, pp. 523–530.

de Brebisson, A., Montana, G., 2015. Deep neural networks for anatomical brain segmentation. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on. pp. 20–28.

Doubal, F. N., MacLullich, A. M. J., Ferguson, K. J., Dennis, M. S., Wardlaw, J. M., Mar 2010. Enlarged perivascular spaces on mri are a feature of cerebral small vessel disease. Stroke 41 (3), 450–454.

Farrell, C., Chappell, F., Armitage, P. A., Keston, P., MacLullich, A., Shenkin, S., Wardlaw, J. M., aug 2008. Development and initial testing of normal reference MR images for the brain at ages 65–70 and 75–80 years. European Radiology 19 (1), 177–183.

Fazekas, F., Chawluk, J. B., Alavi, A., Hurtig, H. I., Zimmerman, R. A., 1987. Mr signal abnormalities at 1.5 t in alzheimer's dementia and normal aging. American Journal of Neuroradiology 8 (3), 421–6.

Feis, D.-L., Brodersen, K. H., von Cramon, D. Y., Luders, E., Tittgemeyer, M., apr 2013. Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. NeuroImage 70, 250–257.

González-Castro, V., del C. Valdés Hernández, M., Armitage, P. A., Wardlaw, J. M., 2016. Texture-based classification for the automatic rating of the perivascular spaces in brain MRI. Procedia Computer Science 90, 9 – 14, 20th Conference on Medical Image Understanding and Analysis (MIUA 2016).

González-Castro, V., Valdés Hernández, M. d. C., Armitage, P. A., Wardlaw, J. M., 2016. Automatic rating of perivascular spaces in brain MRI using bag of visual words. In: Image Analysis and Recognition: 13th International Conference, ICIAR 2016, Proceedings. Springer International Publishing, pp. 642–649.

Haralick, R. M., Shanmugam, K., Dinstein, I., Nov. 1973. Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics SMC-3 (6), 610–621.

Ithapu, V., Singh, V., Lindner, C., Austin, B. P., Hinrichs, C., Carlsson, C. M., Bendlin, B. B., Johnson, S. C., 2014. Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies. Human brain mapping 35 (8), 4219–4235.

Laitinen, L. V., Chudy, D., Tengvar, M., Hariz, M. I., Bergenheim, A. T., Nov 2000. Dilated perivascular spaces in the putamen and pallidum in patients with parkinson's disease scheduled for pallidotomy: a comparison between mri findings and clinical symptoms and signs. Mov Disord 15 (6), 1139–1144.

Lutkenhoff, E. S., Rosenberg, M., Chiang, J., Zhang, K., Pickard, J. D., Owen, A. M., Monti, M. M., 2014. Optimized brain extraction for pathological brains (optiBET). PLoS One 9 (12), e115551.

MacLullich, A. M. J., nov 2004. Enlarged perivascular spaces are associated with cognitive function in healthy elderly men. Journal of Neurology, Neurosurgery & Psychiatry 75 (11), 1519–1523.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. University of California Press, Berkeley, Calif., pp. 281–297.

Munsell, B. C., Wee, C.-Y., Keller, S. S., Weber, B., Elger, C., da Silva, L. A. T., Nesland, T., Styner, M., Shen, D., Bonilha, L., 2015. Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. NeuroImage 118, 219 – 230.

Nam, K. W., Castellanos, N., Simmons, A., Froudist-Walsh, S., Allin, M. P., Walshe, M., Murray, R. M., Evans, A., Muehlboeck, J.-S., Nosarti, C., jul 2015. Alterations in cortical thickness development in preterm-born individuals: Implications for high-order cognitive functions. NeuroImage 115, 64–75.

Ojala, T., Pietikainen, M., Maenpaa, T., Jul. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7), 971–987.

Pantoni, L., 2010. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. Lancet Neurol 9 (7), 689–701.

Patankar, T. F., Baldwin, R., Mitra, D., Jeffries, S., Sutcliffe, C., Burns, A., Jackson, A., jan 2007. Virchow–robin space dilatation may predict resistance to antidepressant monotherapy in elderly patients with depression. Journal of Affective Disorders 97 (1-3), 265–270.

Potter, G. M., Chappell, F. M., Morris, Z., Wardlaw, J. M., 2015a. Cerebral perivascular spaces visible on magnetic resonance imaging: development

of a qualitative rating scale and its observer reliability. Cerebrovasc Dis 39 (3-4), 224–231.

Potter, G. M., Doubal, F. N., Jackson, C. A., Chappell, F. M., Sudlow, C. L., Dennis, M. S., Wardlaw, J. M., 2015b. Enlarged perivascular spaces and cerebral small vessel disease. Int J Stroke 10 (3), 376–381.

Ramirez, J., Berezuk, C., McNeely, A. A., Scott, C. J. M., Gao, F., Black, S. E., 2015. Visible Virchow-Robin spaces on magnetic resonance imaging of Alzheimer's disease patients and normal elderly from the Sunnybrook Dementia Study. J Alzheimers Dis 43 (2), 415–424.

Roy, P. K., Bhuiyan, A., Janke, A., Desmond, P. M., Wong, T. Y., Abhayaratna, W. P., Storey, E., Ramamohanarao, K., 2015. Automatic white matter lesion segmentation using contrast enhanced FLAIR intensity and markov random field. Computerized Medical Imaging and Graphics 45, 102 – 111.

Schölkopf, B., Smola, A. J., 2001. Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press.

Sivic, J., Zisserman, A., 2003. Video google: a text retrieval approach to object matching in videos. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. pp. 1470–1477 vol.2.

Staals, J., Booth, T., Morris, Z., Bastin, M. E., Gow, A. J., Corley, J., Redmond, P., Starr, J. M., Deary, I. J., Wardlaw, J. M., 2015. Total MRI load of cerebral small vessel disease and cognitive ability in older people. Neurobiol Aging 36 (10), 2806–2811.

Staals, J., Makin, S. D. J., Doubal, F. N., Dennis, M. S., Wardlaw, J. M., 2014. Stroke subtype, vascular risk factors, and total MRI brain small-vessel disease burden. Neurology 83 (14), 1228–1234.

Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J. V., Rueckert, D., jul 2014. Multiple instance learning for classification of dementia in brain MRI. Medical Image Analysis 18 (5), 808–818.

Valdés Hernández, M. d. C., Armitage, P. A., Thrippleton, M. J., Chappell, F., Sandeman, E., Muñoz Maniega, S., Shuler, K., Wardlaw, J. M., 2015. Rationale, design and methodology of the image analysis protocol

for studies of patients with cerebral small vessel disease and mild stroke. Brain Behav 5 (12), e00415.

Valdés Hernández, M. d. C., Piper, R. J., Wang, X., Deary, I. J., Wardlaw, J. M., 2013. Towards the automatic computational assessment of enlarged perivascular spaces on brain magnetic resonance images: a systematic review. J Magn Reson Imaging 38 (4), 774–785.

Vapnik, V., 1995. The Nature of Statistical Learning Theory, 2nd Edition. Springer.

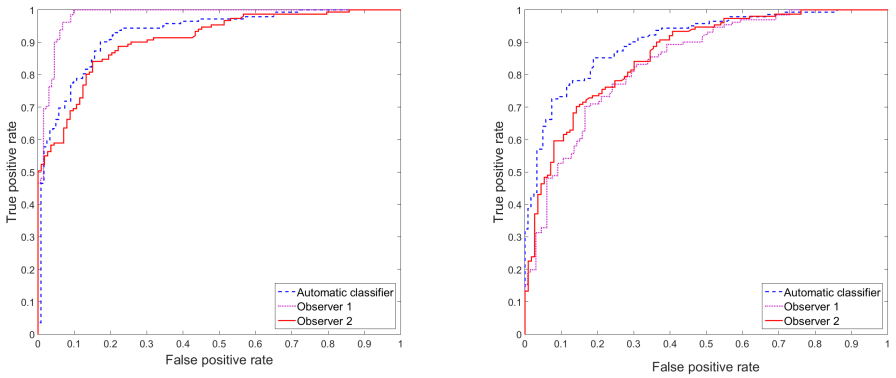
Wang, X., Valdés Hernández, M. D. C., Doubal, F., Chappell, F. M., Piper, R. J., Deary, I. J., Wardlaw, J. M., 2016. Development and initial evaluation of a semi-automatic approach to assess perivascular spaces on conventional magnetic resonance images. J Neurosci Methods 257, 34–44.

Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., et al, 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol 12 (8), 822–838.

Wuerfel, J., Haertle, M., Waiczies, H., Tysiak, E., Bechmann, I., Wernecke, K. D., Zipp, F., Paul, F., aug 2008. Perivascular spaces–MRI marker of inflammatory activity in the brain? Brain 131 (9), 2332–2340.

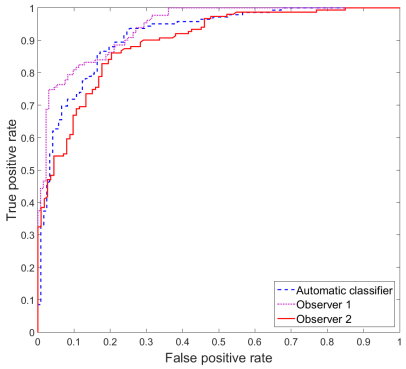
Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging 20 (1), 45–57.

Zuiderveld, K., 1994. Contrast limited adaptive histogram equalization. In: Graphics gems IV. Academic Press Professional, Inc., pp. 474–485.



(a)

(b)



(c)

Figure 10: ROC curves showing the performance for each outcome variable in the regression models 1, 2 and 3 ((a), (b) and (c) respectively). In the model 1 (a) the AUCs of the classifier, Observer 1 and Observer 2 were 0.9265, 0.9813 and 0.9074, respectively. In the model 2 (b) the AUCs of the classifier, Observer 1 and Observer 2 were 0.9041, 0.8395 and 0.8622, respectively. In the model 3 (c) the AUCs of the classifier, Observer 1 and Observer 2 were 0.9152, 0.9411 and 0.8934, respectively